

Your attention is desired



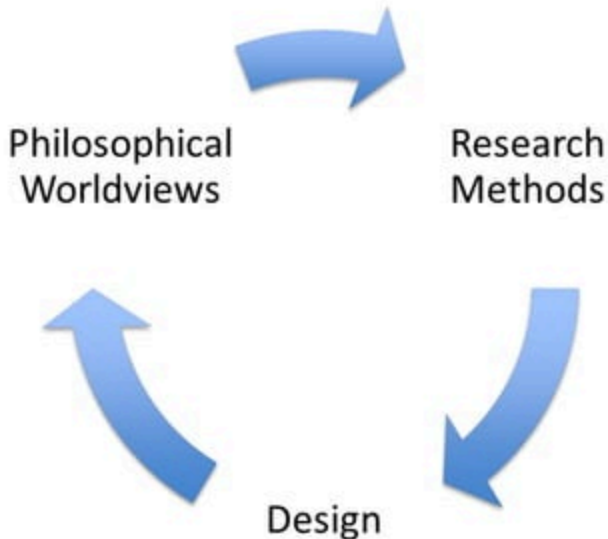
Research Approach: An Overview

Research Approach: Concept

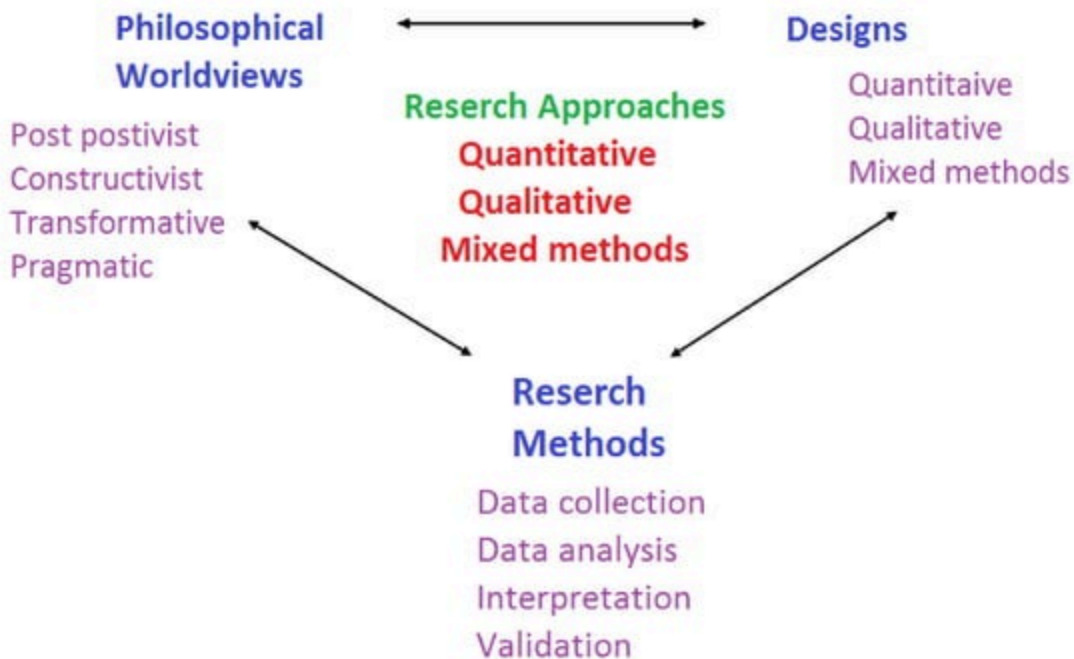
Plans and the procedure for research that span the steps from broad assumptions to detailed methods of data collection, analysis, and interpretation.

The overall decision involves which approach should be used to study a topic. Informing this decision should be the philosophical assumptions the researcher brings to the study; procedures of inquiry (called research designs); and specific research methods of data collection, analysis, and interpretation.

Components of Research Approach



Interconnection



A basic set of Beliefs that guide Actions

source: Guba, 1990,p.17, cited in Creswell, 2014, P.6

Post positivism	Constructivism
<ul style="list-style-type: none">• Determination• Reductionism• Empirical observation and measurement• Theory verification	<ul style="list-style-type: none">• Understanding• Multiple participant meanings• Social and historical construction• Theory generation
Transformative	Pragmatism
<ul style="list-style-type: none">• Political• Power and justice oriented• Collaborative• Change-oriented	<ul style="list-style-type: none">• Consequences of actions• Problem-centered• Pluralistic• Real-world practice oriented

Postpositivist Worldview

Positivist:

- Traditional form of research
- Scientific method
- Empirical science

Postpositivist:

- Challenges the notion of absolute truth
- Deterministic philosophy in which causes determine effects



Postpositivist Worldview

- Assesses causes that influence outcomes
- Reduce ideas to a small set of ideas to test out as variables
- Research is governed by research questions or hypothesis
- Careful observation and measurement of observations
- Testing a theory



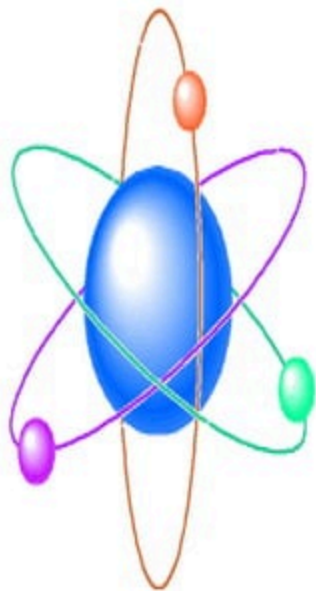
5 Principals of Post positivism WV

1. Knowledge is conjectural...
 - Cannot prove a hypothesis; we state failure to reject the hypothesis or “accept it”
1. Research is the process of making claims...
 - Start with deductively testing a theory
1. Data, evidence, and rational considerations shape knowledge
 - Collecting information through the use of instruments



5 Principals of Post positivism WV (Contd.)

4. Research seeks to develop relevant, true statements...
 - Advances the understanding of relationship among variables
4. Being objective is essential...
 - Validity and reliability are critical



Constructivist Worldview

- Belief that individuals seek understanding of the world
- Belief that people develop subjective meanings of their experiences
- Researchers look for the complexity of views, not the reduction of variables that explain maximal variance
- Interview questions are often very open ended to allow participants to construct their own knowledge of situations



Constructivist Worldview (contd.)

- Researchers focus on the processes and interactions
- Recognition that researcher's personal background shape interpretation and meaning (interpretivism)
- Develops inductively a theory or pattern of meaning instead of testing from theory (grounded approach)



3 Principals of Constructivism WV

1. Construct meaning through interacting with the world we interpret
 - Researches use open ended questions
1. Make sense of the world based on personal historical and societal perspectives
 - Interpretation is also tied to researcher
1. Generation of meaning is social from and within community
 - Meaning generated from data collected



Transformative Worldview

- Developed from a belief that research methods do not fit marginalized individuals...
- Not well defined, includes researchers from various foci and includes: feminist, critical theorists, racial and ethnic minorities, persons with disabilities
- Individuals often overlooked



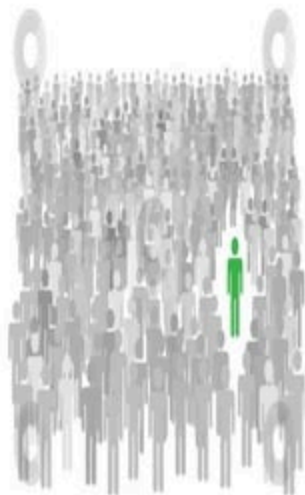
Transformative Worldview (Contd.)

- Element of advocacy in research; the research is directional towards a cause
- Research contains an action agenda that may change the lives of those involved (differs from action research)
- Issues address important aspects: empowerment, inequality, and oppression...



4 Principals of Transformative WV

1. Research focuses on lives of people who have traditionally been marginalized
2. Research focuses on inequities based on gender, race, ethnicity, and disability...
3. Links to political and social action to inequalities
4. Utilize program theory and why problems of oppression, domination and power exist



Pragmatism

- Arises from actions, situation and consequences
- Focus on “what works”
- Focus on problem more than methods
- Practical more than focus on methods



8 Principals of Pragmatism WV

1. No commitment to one philosophy or method
2. Researchers have freedom to chose methods, techniques etc...
3. Look at many ways of collecting data and comparing it (triangulation)
4. "Truth is what works at that time"



8 Principals of Pragmatism WV (Contd.)

5. There is an intended purpose for the research and an examination of consequences
6. Research is situated in context (social, historical, etc.)
7. Practical purpose for research; no need to focus on the nature of the world
8. Very open approach to collecting, analyzing and looking at data



Three Approaches

- **Quantitative** (Positivism and Post positivism)
- **Qualitative** (Constructivism & Transformative)
- **Mixed Methods** (Pragmatism)
- **Logical Theoretical Research**
- **Participatory Research**

Qualitative and quantitative approaches should not be viewed as rigid, distinct categories, polar opposites, or dichotomies. Instead, they represent different ends on a continuum (Newman & Benz, 1998)



Comparative View of Approaches

	Quantitative	Mixed methods	Qualitative
Scientific method	Deductive or 'top down', tests hypotheses and theory with data	Deductive and inductive	Inductive or 'bottoms up', generates new hypotheses and grounded theory from data obtained
View of human behavior	Behavior is regular and predictable	Behavior is somewhat predictable	Behavior is fluidic
research objectives	Description, explanation and prediction	Multiple objectives	Discovery, exploration and discovery
Focus	Narrow-angle lens, testing specific hypotheses	Multi lens focus	Wide-angle and deep angle lens, examining breadth and depth of the phenomenon

Comparative View of Approaches

	Quantitative	Mixed methods	Qualitative
Nature of Observation	Study of behavior under controlled conditions	Study of behavior in more than one context and conditions	Study of behavior in natural environment Study of context or conditions in which behavior occurred
Nature of reality	Objective (different observers agree on what is observed)	Commonsense realism and pragmatic view of world (what works is real or true)	Subjective, personal and socially constructed
Data	Measurements using precise and validated instruments (physical instruments, close ended questions, rating scales etc.)	Multiple forms	Researcher is primary data collection instrument, using in-depth interview, observations, open ended questions etc.

Comparative View of Approaches

	Quantitative	Mixed methods	Qualitative
Nature of data	Variables	Mix of variables, words, images	Words, images, categories, patterns
Data analysis	Statistical relationships	Quantitative symptoms and qualitative support	Search for patterns, themes and holistic features
Results	Generalizing	Corroborated findings may generalize	Particularistic findings Representation of insider i.e. 'emic' view point
Final report form	Statistical report	Eclectic and pragmatic	Narrative even with direct quotations of research participants

Comparative View of Approaches

	Quantitative	Mixed methods	Qualitative
Typology of designs	<p>Experimental and Non experimental Designs</p> <p>Experimental designs include: True Experiments (Randomized control trials)</p> <p>Quasi Experiments (Not as random, more natural)</p> <p>Causal Comparative (Explain variation, regression based)</p> <p>Correlation based (Relation among two or more variables)</p> <p>Longitudinal Analysis (Repeated measures of more than three points)</p>	<p>Mixed methods, Multi methods and Mixed models</p> <p>Interdisciplinary and multidisciplinary</p> <p>Convergent Parallel (Mixing occurs at the end)</p> <p>Explanatory Sequential mixed method (Qualitative followed by Qualitative phase)</p> <p>Exploratory Sequential mixed method (Reverse)</p> <p>Transformative mixed methods (Draws from ideas of social justice)</p>	<p>Phenomenology, Ethnography, Case study, Grounded theory, Narrative, Self Inquiry, Cognitive interviews, Iterative designs, Historical</p>

Comparative View of Approaches

Quantitative	Mixed methods	Qualitative
Non Experimental methods include: Descriptive, correlational, Ex-post facto, Developmental, Epidemiological, Survey , Methodological, Meta analysis, Secondary data analysis, Outcome research, Evaluation studies, Operational research	Embedded mixed methods (Embedding Quantitative or Qualitative within the other) Multiphase mixed methods (Over time multiple parts)	

Logical, theoretical research

By a *logical theoretical* research approach is meant formal deduction of logical consequences from a set of initial assumptions (axioms). If the axioms are true and the rules are logically sound, the consequences are true as well.

According to Hirschheim et al. (1995),

"data modeling is first and foremost a social and organizational activity and very little, if anything (except consulting folklore) is known how data modeling is exercised in practice and what its impacts are on organizations, their information systems management, and business operations."

Participatory action research

Reason, 1994; van Meel, 1993 refer participatory research as a set of methods to research on social systems in which the researcher actively engage in the process under investigation (the actors of the social system being studied can be considered as co-researchers). Meel exemplifies this approach to research: First, an initial case study is performed for identification of problems, followed by theory development and implementation of a prototype. Then Prototype is employed to full-scale project, researcher participates and reflects upon the use of the prototype with the actors.

Walsham (1995) points out the problems of being perceived to have a personal stake in the researched project, and reporting on one's own role within the project as challenging problems of participatory action research.

STATISTICS IN
RESEARCH (PART – 1
DESCRIPTIVE)



Role of statistics in research

- Designing research
- Analyzing data
- Draw conclusion about research

Two major areas of statistics

□ Descriptive statistics

- It concern with development of certain indices from the raw data.
- It summarizes collected/ classified data.

□ Inferential statistics

- It adopts the process of generalization from small groups (i.e., samples) to population.
- It also known as sampling statistics.
- It concern with two major problems
 - Estimation of population parameters.
 - Testing of statistical hypothesis.

Descriptive statistics

- Measure of central tendency (Statistical Average)
- Measure of dispersion
- Measure of asymmetry (Skewness)
- Measure of relationship
- Other measures

Measure of Central tendency

- It also known as statistical average.
- Mean, Median and Mode are the popular averages.
- Geometric and Harmonic mean are also sometime used.

Mean

- It also known as arithmetic average.
- The mean is found by adding all the values in the set, then dividing the sum by the number of

$$\text{Mean (or } \bar{X})^* = \frac{\sum X_i}{n} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

where \bar{X} = The symbol we use for mean (pronounced as X bar)

\sum = Symbol for summation

X_i = Value of the i th item X , $i = 1, 2, \dots, n$

n = total number of items

Example

- Find the mean of following numbers 8, 9, 10, 10, 10, 11, 11, 11, 12, 13

The mean is the usual average:

- $n = 10$
- $(8 + 9 + 10 + 10 + 10 + 11 + 11 + 11 + 12 + 13) \div 10 = 105 \div 10 = 10.5$

Weighted Mean

- The weighted average formula is used to calculate the average value of a particular set of numbers with different levels of weight.

$$\bar{X}_w = \frac{\sum w_i X_i}{\sum w_i}$$

where \bar{X}_w = Weighted item
 w_i = weight of i th item X
 X_i = value of the i th item X

Example

- Final marks has distribution of internal marks, Mid semester marks and End semester marks. Each one carries following weightage

Internal	Mid-Sem	End-Sem
30	20	50

Shakthi got 80 marks in internal, 75 marks in Mid sem and 89 marks in End sem. Then what is his final mark? Ans = 83.5

Frequency distribution Mean

$$\bar{X} = \frac{\sum f_i X_i}{\sum f_i} = \frac{f_1 X_1 + f_2 X_2 + \dots + f_n X_n}{f_1 + f_2 + \dots + f_n = n}$$

Isabella went up and down the street to find out how many parking spaces each house had. Here are her results: What is the mean number of Parking Spaces?

Parking Spaces	Frequency
1	15
2	27
3	8
	5

$$\text{Mean} = \frac{15 \times 1 + 27 \times 2 + 8 \times 3 + 5 \times 4}{15 + 27 + 8 + 5}$$

$$\text{Mean} = 2.05$$

Median

- *Median* is the value of the middle item of series when it is arranged in ascending or descending order of magnitude.
- It divides the series into two halves; in one half all items are less than median, whereas in the other half all items have values higher than median

$$\text{Median } (M) = \text{Value of } \left(\frac{n+1}{2}\right)\text{th item}$$

Example

- 21, 18, 24, 19, 27
- Arrange the number in ascending order
 - 18, 19, 21, 24, 27 (21 is the median)
- If there are two middle numbers,
 - 18, 19, 21, 25, 27, 28
 - $(21 + 25) / 2 = 23$

Mode

- The number that appears most frequently in a set of numbers.
- It is useful in the studies of popular. (Popular shoe size, popular cap, most demanded product etc.,)
- Arrange the numbers in order from least to greatest.
 - 21, 18, 24, 19, 18
- Find the number that is repeated the most.
 - 18, 18, 19, 21, 24 (18 is the mode).

Geometric Mean

- It is defined as the n th root of the product of the values of n times in a given series.
- The most frequently used application of this average is in the determination of average percent of change.

$$\begin{aligned}\text{Geometric mean (or G.M.)} &= \sqrt[n]{\pi X_i} \\ &= \sqrt[n]{X_1 \cdot X_2 \cdot X_3 \dots X_n}\end{aligned}$$

where

G.M. = geometric mean,

n = number of items.

X_i = i th value of the variable X

π = conventional product notation

Example

- What is the geometric mean of 4,6 and 9 ?

Ans: 6

- Your investment earns 20% during the first year, but then realizes a loss of 10% in year 2, and another 10% in year 3
 - Calculate a growth factor for each year. $(1+.2)$ for Year 1, $(1-.1)$ for year 2 and $(1-.1)$ for Year 3
 - Multiply the 3 growth factors and take the 3rd root.
 - Thus, geometric mean = $0.990578-1=-0.009422$.
 - So your investment losing roughly .9 percent of every year.

Harmonic Mean

- It is defined as reciprocal of the average of reciprocal of values of items of a series.
- It has limited application particularly in time and rate are involved.
- It gives largest weight to smallest value and smallest weight to largest value.

Harmonic Mean

$$\begin{aligned}\text{Harmonic mean (H. M.)} &= \text{Rec.} \frac{\sum \text{Rec} X_i}{n} \\ &= \text{Rec.} \frac{\text{Rec.} X_1 + \text{Rec.} X_2 + \dots + \text{Rec.} X_n}{n}\end{aligned}$$

What is the Harmonic mean of 4,5 and 10?

Ans: 5.45

Example

- For example, suppose that you have four 10 km segments to your automobile trip. You drive your car:
 - 100 km/hr for the first 10 km
 - 110 km/hr for the second 10 km
 - 90 km/hr for the third 10 km
 - 120 km/hr for the fourth 10 km.
- Ans: 103.8 km/hr.

Measure of Dispersion

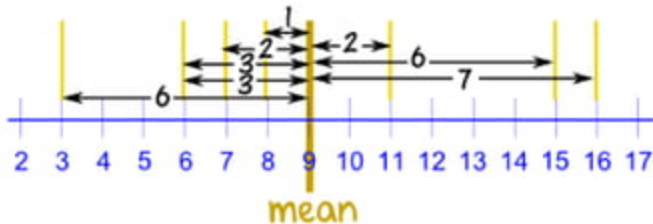
- Average can't reveal the entire story of study. It fail to give ideas about data which distributed around average.
- Measures of dispersion measure how spread out a set of data is.
- Important measure of dispersion are
 - Range
 - Mean Deviation
 - Standard deviation

Range

- It is difference between extreme values.
- Range = Highest value – Lowest value
- What is the range for following data?
- 2, 3, 1, 1, 0, 5, 3, 1, 2, 7, 4, 0, 2, 1, 2, 1, 6, 3, 2, 0, 0, 7, 4, 2, 1, 1, 2, 1, 3, 5, 12
- Range = 12
- Limitation: Its value is never stable, being based on only two values of the variable.

Mean Deviation

- It is the average of difference of the values of items from some average of the series
- Find the mean deviation of 3, 6, 6, 7, 8, 11, 15, 16
 - Find the **mean**
 - Find the absolute distance from the mean.
 - Find mean of those Distances.
 - Ans= 3.75



Coefficient of mean deviation

- When mean deviation is divided by the average used in finding out the mean deviation itself, the resulting quantity is described as the *coefficient of mean deviation*.
- Coefficient of mean deviation is a relative measure of dispersion and is comparable to similar measure of other series.

Standard deviation

- It is most widely used measure of dispersion of a series and is commonly denoted by the symbol ' σ ' (pronounced as sigma).

$$\text{Standard deviation}^* (\sigma) = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n}}$$

Standard deviation

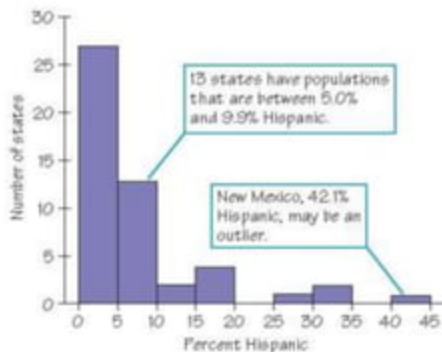
- Find the standard deviation of 53,61,49,67, 55,63.
 - Steps:
 - Find the mean (58)
 - Find the deviation from mean, square it and sum it (230)
 - Divide the above answer by sample size (38.333)

Other terms related to σ

- Variance : Square the standard deviation
- Coefficient of standard deviation: Divide the standard deviation by Mean of that data.
- Coefficient of variation: Multiply the coefficient of standard deviation with 100.

Distribution

- The pattern of outcomes of a variable; it tells us what values the variable takes and how often it takes these values.
- The distribution of data can find with help of histogram.
- Histogram is bar chart.



Steps in making Distribution

- Choose the classes by dividing the range of data into classes of equal width (individuals fit into one class).
- Count the individuals in each class (this is the height of the bar).
- Draw the histogram:
 - The horizontal axis is marked off into equal class widths.
 - The vertical axis contains the scale of counts (frequency of occurrences) for each class

Histogram

- The number of days of Maria's last 15 vacations are listed below. Use the data to make a frequency table with intervals.

4, 8, 6, 7, 5, 4, 10, 6, 7, 14, 12, 8, 10, 15, 12.

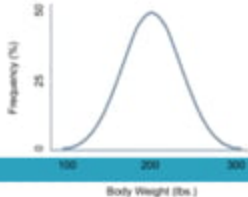
- Step 1: Identify the least and greatest values.
- Step 2: Divide the data into equal intervals.

Histogram

- Step 3: List the intervals in the first column of the table. Count the number of data values in each interval and list the count in the last column. Give the table a title.

Interval	Frequency
4 – 6	5
7 – 9	4
10 – 12	4
13 – 15	2

Normal Distribution



- The distribution of data happens to be perfectly symmetrical.
- It is perfectly bell shaped curve in which case the value of mean \bar{X} = median M = mode Z .

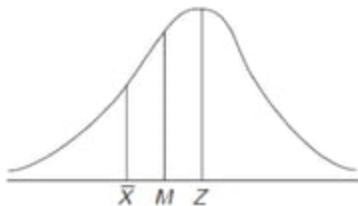
Skewness

- if the curve is distorted (whether on the right side or on the left side), we have asymmetrical distribution which indicates that there is skewness.
- If the curve is distorted on the right side, we have positive skewness.
- If the curve is distorted on the left side, we have negative skewness.

Measure of Skewness



Curve showing positive skewness
In case of positive skewness we have:
 $Z < M < \bar{X}$



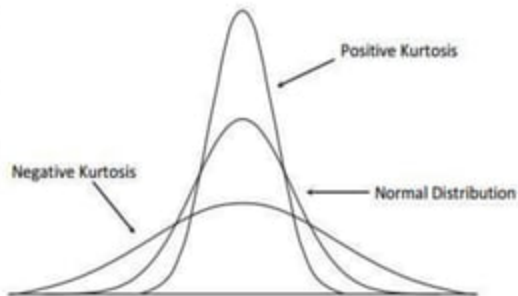
Curve showing negative skewness
In case of negative skewness we have:
 $\bar{X} < M < Z$

Measure of Skewness

- Skewness = $\bar{X} - Z$ or $3(\bar{X} - M)$
- Coefficient of Skewness = $\frac{(\bar{X} - Z)}{\sigma}$
- If skewness value is positive/ negative/ zero, then data are positively/negatively skewed/symmetry.

Kurtosis

- Kurtosis is the measure of flat-toppedness or peakedness of a curve.
- Normal curve is MesoKurtic.
- Positive kurtosis is leptokurt
- Negative kurtosis is platykur



Measure of Relationship

- Univariate Analysis: The analysis is carried out with the description of a single variable.
- Bivariate Analysis: The analysis of two variables simultaneously.
- Multivariate Analysis: The analysis of multiple variables simultaneously.

Measure of Relationship

□ Correlation

- The word Correlation is made of **Co-** (meaning "together"), and **Relation**.
- It answers that Does there exist association or correlation between the two (or more) variables ?

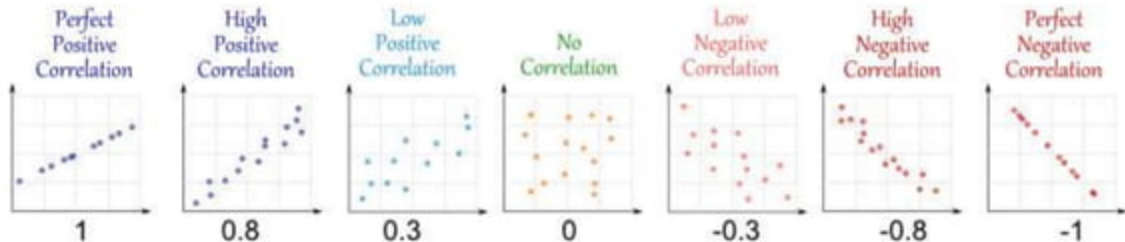
□ Regression

- It answers that:
- Is there any cause and effect relationship between the two variables in case of the bivariate population or between one variable on one side and two or more variables on the other side in case of multivariate population? If yes, of what degree and in which direction?

Karl Pearson's coefficient of correlation

- It is simple correlation and most widely used method. Denotes by r .
- It is also known as the product moment correlation coefficient.
- The value of r lies between ± 1 .

r value and interpretation



We can also say that for a unit change in independent variable, if there happens to be a constant change in the dependent variable in the same direction, then correlation will be termed as perfect positive.

Examples

- **Positive relationships**

- water consumption and temperature.
- study time and grades.

- **Negative relationships:**

- alcohol consumption and driving ability.
- Price & quantity demanded

Karl Pearson Correlation coefficient formula

$$r = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{n \cdot \sigma_x \cdot \sigma_y}$$

If substitute the value of sigma and derive, you will get below equation

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Example

- A researcher want to know the relation between advertisement expenditure and total sales. He took a sample data of 7 companies for one ye

Advertisement exp (Million dollars)	Annual Sales (Million dollars)
9	19
7	13
5	12
8	16
6	15
3	10
4	8

Solution

x	y	xy	x^2	y^2
9	19	171	81	361
7	13	91	49	169
5	12	60	25	144
8	16	128	64	256
6	15	90	36	225
3	10	30	9	100
4	8	32	16	64
42	93	602	280	1319

Solution

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$r = \frac{7(602) - (42)(93)}{\sqrt{[7(280) - (42)^2][7(1319) - (93)^2]}}$$

$$r = \frac{4214 - 3906}{\sqrt{[1960 - 1764][9233 - 8649]}}$$

$$r = \frac{308}{\sqrt{[196][584]}}$$

$$r = \frac{308}{\sqrt{114464}}$$

$$r = \frac{308}{338.33}$$

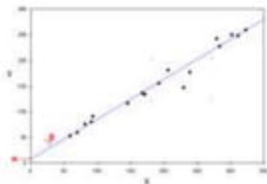
$$r = 0.910$$

Regression

- It is the study of the relationship between variables.
- It also used for prediction of dependent variable.
- Regression types
 - Simple Regression: single explanatory/independent variable

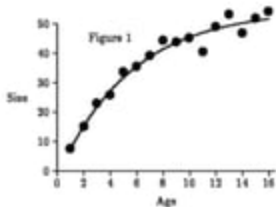
Regression Analysis

- Linear Regression: Straight-line relationship.
- Non-linear: Implies curved relationships.
 - Regression is nothing but try to find out the equation of line/curve.



Linear

→ $y=mx+b$



Non linear

Linear Regression Equation

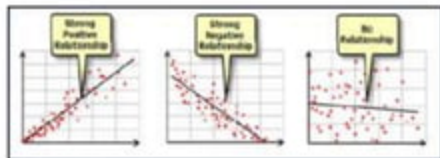
$$\hat{Y} = a + bX$$

Unknown parameter a and b can be calculated by below formula
m = b = slope of line, a = c = intercept

$$b = \Sigma [(x_i - \bar{x})(y_i - \bar{y})] / \Sigma [(x_i - \bar{x})^2]$$

$$a = \bar{y} - b * \bar{x}$$

Above formula gives best fit line by using least square method.



Example for regression

- In the table below, the X column shows scores on the aptitude test. Similarly, the Y column shows statistics grades.

Student	X	Y
1	95	85
2	85	95
3	80	70
4	70	65
5	60	70

Regression

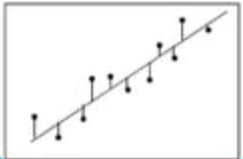
S.no	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	95	85	17	8	289	64	136
2	85	95	7	18	49	324	126
3	80	70	2	-7	4	49	-14
4	70	65	-8	-12	64	144	96
5	60	70	-18	-7	324	49	126
Sum	390	385			730	630	470
Mean	78	77					

$$b = 470/730 = 0.644$$

$$a = 26.768$$

$$\hat{y} = 26.768 + 0.644x$$

Standard error



- It provide an overall measure of how well the model fits the data.
- It represents the average distance that the observed values fall from the regression line.
- Smaller values are better because it indicates that the observations are closer to the fitted line.

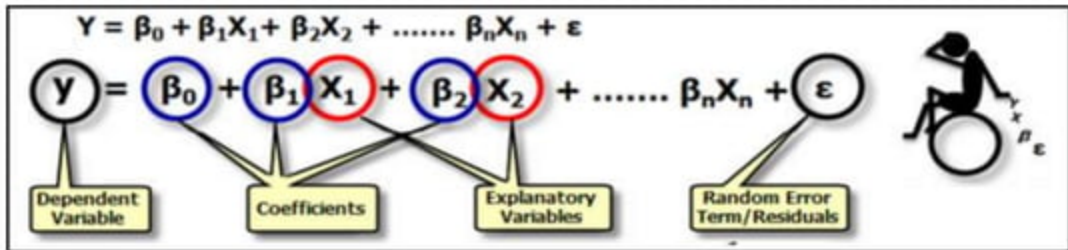
$$\sigma_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N}}$$

Standard error

Y	\bar{Y}	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$
85	81.508	3.492	12.19406
95	87.948	7.052	49.7307
70	71.848	-1.848	3.415104
65	68.628	-3.628	13.16238
70	71.848	-1.848	3.415104
Sum			81.91736

Standard error = 4.0475

Multiple regression



Multicollinearity: Independent variables in multiple regression are highly correlated.

Partial Correlation

- In simple correlation, we measure the strength of the linear relationship between two variables, without taking into consideration the fact that both these variables may be influenced by a third variable.
 - For Ex: when we study the correlation between price and demand, we completely ignore the effect of other factors like money supply, import and exports etc. which definitely have a bearing on the price.
- The correlation co-efficient between two variables X1 and X2, studied partially after eliminating the influence of the third variable X3 from both of them, is the partial correlation co-efficient.

$$r_{x_1 x_2} = \frac{r_{x_1 x_2} - r_{x_1 x_3} \cdot r_{x_2 x_3}}{\sqrt{1 - r_{x_1 x_3}^2} \sqrt{1 - r_{x_2 x_3}^2}}$$

Other Measures

- Index number

- Indicator of average percentage change in a series of figures where one figure (called the base) is assigned an arbitrary value of 100, and other figures are adjusted in proportion to the base.

- Time series analysis

- Unlike the analyses of random samples of observations that are discussed in the context of most other statistics, the analysis of time series is based on the assumption that successive values in the data file represent consecutive measurements taken at equally spaced time intervals.

TESTING OF HYPOTHESIS

Parameter and Statistics

- A measure calculated from population data is called **Parameter**.
- A measure calculated from sample data is called **Statistic**.

	Parameter	Statistic
Size	N	n
Mean	μ	\bar{x}
Standard deviation	σ	s
Proportion	P	p
Correlation coefficient	ρ	r

TESTING OF HYPOTHESIS

Statistical Hypothesis

A Statistical hypothesis is an assumption or any logical statement about the parameter of the population.

E.g.

- India will score on an average 300 runs in the next ODI series.
- The average marks obtained by students at Guj. Uni. in Statistics is atleast 80.
- Proportion of diabetic patients in Gujarat is not more than 10 %
- Students of Guj. Uni. score better than students from other universities

Null hypothesis

A statistical hypothesis which is written for the possible acceptance is called **Null hypothesis**. It is denoted by H_0 .

- In Null hypothesis if the parameter assumes specific value then it is called **Simple hypothesis**.

E.g. $\mu = 280, P=0.10$

- In Null hypothesis if the parameter assumes set of values then it is called **Composite hypothesis**.

E.g. $\mu \geq 280, P \leq 0.10$

Alternative Hypothesis

A statistical hypothesis which is complementary to the Null hypothesis is called **Alternative hypothesis**. It is denoted by H_1 .

Null hypothesis	Alternative hypothesis	
$\mu = \mu_0$	$\mu \neq \mu_0$	Two tailed test
	$\mu > \mu_0$	Right sided one tailed test
	$\mu < \mu_0$	Left sided one tailed test

Problem Statement	Null hypothesis (H0)	Alternative hypothesis (H1)
India will score on an average 300 runs in the next ODI series	$\mu = 300$	$\mu \neq 300$
The average marks obtained by students at Guj Uni in Statistics is atleast 80	$\mu = 80$	$\mu < 80$
Proportion of diabetic patients in Gujarat is not more than 10 %	$P = 0.10$	$P > 0.10$
Students of Guj Uni score better than students from other Universities	$\mu_1 = \mu_2$	$\mu_1 > \mu_2$

Testing of Hypothesis

The procedure to decide whether to accept or reject the null hypothesis is called **Testing of hypothesis**.

Type I and Type II Error

- The error of rejecting the true null hypothesis is called Type I error. The probability of type I error is denoted by α .

$$\alpha = \text{Prob [Reject } H_0 / H_0 \text{ is true]}$$

- The error of accepting the false null hypothesis is called Type II error. The probability of type II error is denoted by β .

$$\beta = \text{Prob [Accept } H_0 / H_0 \text{ is false]}$$

Type I and Type II Error

DECISION	Null Hypothesis	
	TRUE	FALSE
ACCEPT	No Error	Type II Error
REJECT	Type I Error	No Error

Level of Significance

The predetermined value of probability of type I error is called **level of significance**. It is denoted by α .

The most commonly used level of significance are 1% or 5%.

Interpretation: 5% level of significance means in 5 out of 100 cases, it is likely to reject a true null hypothesis.

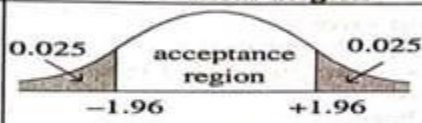
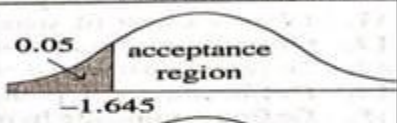

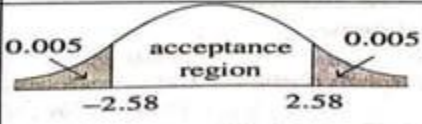
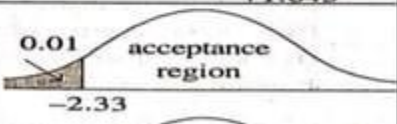

Critical Region

The area of the probability curve corresponding to α is called **critical region**.

i.e. the area under normal curve at which a true null hypothesis is rejected is called area of rejection or critical region.

The remaining region under normal curve is called **acceptance region**.

Critical Region

Level of Significance	Two sided Critical Region	One-sided Critical Region
5%	 <p>0.025 acceptance region 0.025 -1.96 +1.96</p>	 <p>0.05 acceptance region -1.645</p>  <p>acceptance region 0.05 +1.645</p>
1%	 <p>0.005 acceptance region 0.005 -2.58 2.58</p>	 <p>0.01 acceptance region -2.33</p>  <p>acceptance region 0.01 2.33</p>

Power of Test

The probability of rejecting the false null hypothesis is called the **Power of the test.**

It is denoted by $1 - \beta$.

i.e. $1 - \beta = \text{Prob} [\text{Reject } H_0 / H_0 \text{ is false}]$

Test Statistics

If the sample size is more than or equal to 30, it is called a **large sample** and if it is less than 30, it is called a **small sample**.

Different test statistic is used for testing of hypothesis based on the size of the sample.

- For a large sample, test statistic **z** is used.
- For a small sample, test statistic **t** is used.

Steps of Testing of Hypothesis

- Step 1: Setting up Null hypothesis
- Step 2: Setting up Alternative hypothesis
- Step3: Calculating test statistics
- Step 4: Determining table value of test statistics
- Step 5: Conclusion
 - If test statistics \leq table value, Null hypothesis is **Accepted**
 - If test statistics $>$ table value, Null hypothesis is **Rejected**

Large Sample test

- Test of Single Mean
- Test of significance of difference between two means
- Test of significance of difference between two std. deviation
- Test of Single Proportion
- Test of significance of difference between two proportions

z table value

	1 %	5%	10%
Two tailed test (\neq)	2.58	1.96	1.645
One tailed test ($>$ or $<$)	2.33	1.645	1.28

Test 1: Test of Single Mean

- Step 1: Null hypothesis $H_0: \mu = \mu_0$
- Step 2: Alternative hypothesis $H_1: \mu \neq \mu_0$ or $\mu > \mu_0$ or $\mu < \mu_0$
- Step 3: Test statistics

$$z = \frac{|\bar{x} - \mu|}{\frac{\sigma}{\sqrt{n}}} \quad \text{or} \quad \frac{|\bar{x} - \mu|}{\frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}}$$

Finite population correction factor

Denominator is the Standard Error of sample mean i.e. $S.E.(\bar{x})$

- Step 4: Table value of z at α % level of significance
- Step 5: If $z \leq z$ table value, H_0 is Accepted
If $z > z$ table value, H_0 is Rejected

Case Study 1

It is hoped that a newly developed pain reliever will more quickly produce perceptible reduction in pain to patients after minor surgeries than a standard pain reliever.

The standard pain reliever is known to bring relief in an average of 3.5 minutes with standard deviation 2.1 minutes. To test whether the new pain reliever works more quickly than the standard one, 50 patients with minor surgeries were given the new pain reliever and their times to relief were recorded. The experiment yielded sample mean $\bar{x} = 3.1$ minutes and sample standard deviation $s = 1.5$ minutes. Is there sufficient evidence in the sample to indicate, at the 5% level of significance, that the newly developed pain reliever does deliver perceptible relief more quickly?

Case Study 2

A cosmetics company fills its best-selling 8-ounce jars of facial cream by an automatic dispensing machine.

The machine is set to dispense a mean of 8.1 ounces per jar. Uncontrollable factors in the process can shift the mean away from 8.1 and cause either underfill or overfill, both of which are undesirable. In such a case the dispensing machine is stopped and recalibrated. Regardless of the mean amount dispensed, the standard deviation of the amount dispensed always has value 0.22 ounce. A quality control engineer routinely selects 30 jars from the assembly line to check the amounts filled. On one occasion, the sample mean is 8.2 ounces and the sample standard deviation is 0.25ounce. Determine if there is sufficient evidence in the sample to indicate, at the 1% level of significance, that the machine should be recalibrated

Test 2: Test of significance of difference between two means

- Step 1: Null hypothesis $H_0: \mu_1 = \mu_2$
- Step 2: Alternative hypothesis $H_1: \mu_1 \neq \mu_2$ or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$
- Step 3: Test statistics

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Denominator is the Standard Error of difference of sample means i.e. $S.E.(\bar{x}_1 - \bar{x}_2)$

- Step 4: Table value of z at α % level of significance
- Step 5: If $z \leq z$ table value, H_0 is Accepted
If $z > z$ table value, H_0 is Rejected

Case Study 1

A nutritionist is interested in whether two proposed diets, Diet A and Diet B work equally well in providing weight-loss for customers.

In order to assess a difference between the two diets, she puts 50 customers on Diet A and 60 other customers on the Diet B for two weeks. Those on the former had weight losses with an average of 11 pounds and a standard deviation of 3 pounds, while those on the latter lost an average of 8 pounds with a standard deviation of 2 pounds.

Do the diets differ in terms of their weight loss?

Case Study 2

To compare customer satisfaction levels of two competing cable television companies, 174 customers of Company 1 and 355 customers of Company 2 were randomly selected and were asked to rate their cable companies on a five-point scale, with 1 being least satisfied and 5 most satisfied. The survey results are summarized in the following table:

Company 1	Company 2
$n_1 = 174$	$n_2 = 355$
$\bar{x}_1 = 3.51$	$\bar{x}_2 = 3.24$
$s_1 = 0.51$	$s_2 = 0.52$

Test at the 1% level of significance whether the data provide sufficient evidence to conclude that Company 1 has a higher mean satisfaction rating than does Company 2.

Test 3: Test of significance of difference between two std. deviations

- Step 1: Null hypothesis $H_0: \sigma_1 = \sigma_2$
- Step 2: Alternative hypothesis $H_1: \sigma_1 \neq \sigma_2$ or $\sigma_1 > \sigma_2$ or $\sigma_1 < \sigma_2$
- Step 3: Test statistics

$$z = \frac{|s_1 - s_2|}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}}$$

Denominator is the Standard Error of difference of sample standard deviation i.e. S.E.($s_1 - s_2$)

- Step 4: Table value of z at α % level of significance
- Step 5: If $z \leq z$ table value, H_0 is Accepted
If $z > z$ table value, H_0 is Rejected

Case Study

To compare customer satisfaction levels of two competing cable television companies, 174 customers of Company 1 and 355 customers of Company 2 were randomly selected and were asked to rate their cable companies on a five-point scale, with 1 being least satisfied and 5 most satisfied. The survey results are summarized in the following table:

Company 1	Company 2
$n_1 = 174$	$n_2 = 355$
$\bar{x}_1 = 3.51$	$\bar{x}_2 = 3.24$
$s_1 = 0.51$	$s_2 = 0.52$

Test at the 1% level of significance whether the data provide sufficient evidence to conclude that there is significant difference in standard deviation of Company 1 and Company 2.

Test 4: Test of Single Proportion

- Step 1: Null hypothesis $H_0: P = P_0$
- Step 2: Alternative hypothesis $H_1: P \neq P_0$ or $P > P_0$ or $P < P_0$
- Step 3: Test statistics

$$Z = \frac{|p - P|}{\sqrt{\frac{P \cdot Q}{n}}}$$

Denominator is the Standard Error of sample proportion i.e. S.E.(p)

- Step 4: Table value of z at α % level of significance
- Step 5: If $z \leq z$ table value, H_0 is Accepted
If $z > z$ table value, H_0 is Rejected

Case Study 1

Globally the long-term proportion of new-borns who are male is 51.46%. A researcher believes that the proportion of boys at birth changes under severe economic conditions. To test this belief randomly selected birth records of 5,000 babies born during a period of economic recession were examined. It was found in the sample that 52.55% of the new-borns were boys.

Determine whether there is sufficient evidence, at the 10% level of significance, to support the researcher's belief.

Case Study 2

A soft drink maker claims that a majority of adults prefer its leading beverage over that of its main competitor's. To test this claim 500 randomly selected people were given the two beverages in random order to taste. Among them, 270 preferred the soft drink maker's brand, 211 preferred the competitor's brand, and 19 could not make up their minds.

Determine whether there is sufficient evidence, at the 5% level of significance, to support the soft drink maker's claim against the default that the population is evenly split in its preference.

Test 5: Test of significance of difference between two proportions

- Step 1: Null hypothesis $H_0: P_1 = P_2$
- Step 2: Alternative hypothesis $H_1: P_1 \neq P_2$ or $P_1 > P_2$ or $P_1 < P_2$
- Step 3: Test statistics

$$z = \frac{|p_1 - p_2|}{\sqrt{\frac{P_1 \cdot Q_1}{n_1} + \frac{P_2 \cdot Q_2}{n_2}}} \quad \text{or} \quad \frac{|p_1 - p_2|}{\sqrt{\hat{p} \cdot \hat{q} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where } \hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Denominator is the Standard Error of difference in sample proportion i.e. S.E.($p_1 - p_2$)

- Step 4: Table value of z at α % level of significance
- Step 5: If $z \leq z$ table value, H_0 is Accepted
If $z > z$ table value, H_0 is Rejected

Case Study 1

Voters in a particular city who identify themselves with one or the other of two political parties were randomly selected and asked if they favour a proposal to remove article 370 from Kashmir. The results are:

	Party A	Party B
Sample Size	150	200
No. in favour	90	140

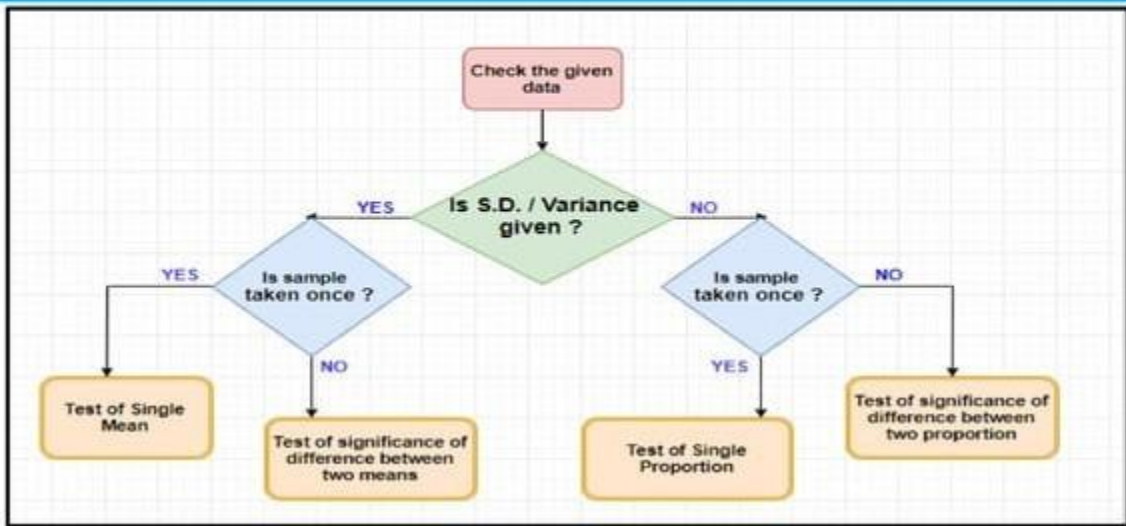
Test, at the 5% level of significance, the hypothesis that the proportion of members of Party A who favour the proposal is less than the proportion of members of Party B who do.

Case Study 2

Suppose the Acme Drug Company develops a new drug, designed to prevent colds. The company states that the drug is equally effective for men and women. To test this claim, they choose a simple random sample of 100 women and 200 men who were suffering from cold.

At the end of the study, 38% of the women were cured from cold; and 51% of the men were cured from cold. Based on these findings, can we reject the company's claim that the drug is equally effective for men and women? Use a 0.05 level of significance.

Flow chart for selecting test



Inferential Statistics

Descriptive & Inferential Statistics

Descriptive Statistics

Organize

- Summarize
- Simplify
- Presentation of data



Describing data

Inferential Statistics

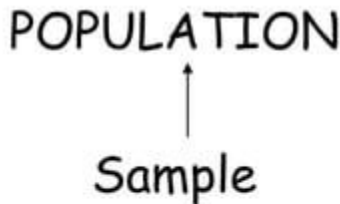
- Generalize from samples to pops
- Hypothesis testing
- Relationships among variables



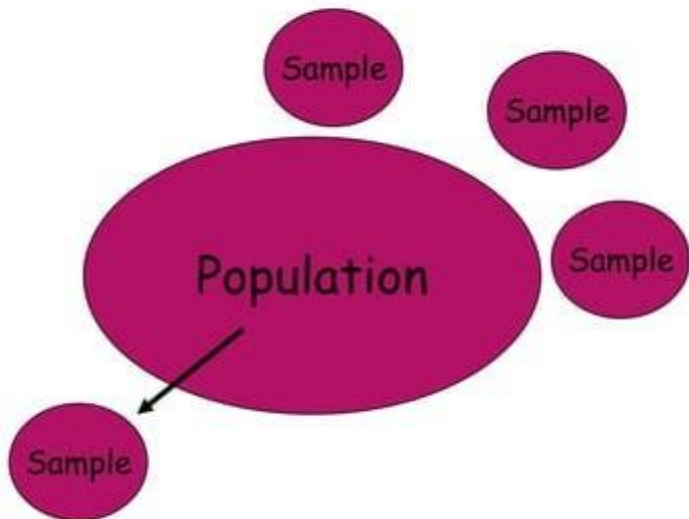
Make predictions

Inferential Statistics

- Inferential statistics are used to draw conclusions about a population by examining the sample

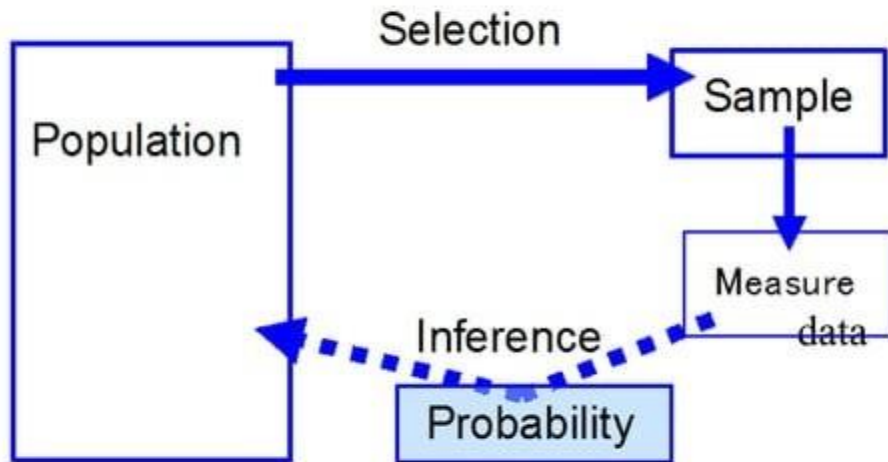


Inferential Statistics



Draw inferences about the larger group

Chain of Reasoning for Inferential Statistics



Are our inferences valid? ... Best we can do is to calculate probability about inferences

Inferential Statistics

- Accuracy of inference depends on **representativeness** of sample from population
- random selection
 - equal chance for anyone to be selected makes sample **more representative**

Inferential Statistics

- Inferential statistics help researchers test hypotheses and answer research questions, and derive meaning from the results

Sampling Error: variability among samples due to chance vs population



Or true differences? Are just due to sampling error?
Probability.....

Error...misleading...not a mistake

Inferential Statistics

- Researchers set the **significance level** for each statistical test they conduct

Alternative and Null Hypotheses

- If the .05 level is achieved (p is equal to or less than .05), then a researcher **rejects the H_0 and accepts the H_1**
- If the the .05 significance level is not achieved, then the H_0 is retained

Degrees of Freedom

- Degrees of freedom (*df*) are the way in which the scientific tradition accounts for **variation due to error**
 - it specifies **how many values vary** within a statistical test
 - scientists recognize that collecting data can never be error-free
 - each piece of data collected can **vary**, or carry error that we cannot account for
 - by including *df* in statistical computations, scientists help account for this error

Inferential Statistics: 5 Steps

- To determine if SAMPLE means come from same population, use 5 steps with inferential statistics

1. State Hypothesis

- H_0 : no difference between 2 means; any difference found is due to sampling error
 - any significant difference found is not a TRUE difference, but CHANCE due to sampling error
- results stated in terms of *probability* that H_0 is false
 - findings are stronger if can reject H_0
 - therefore, need to specify H_0 and H_1

Steps in Inferential Statistics

2. Level of Significance

- Probability that sample means are different enough to reject H_0 (.05 or .01)
 - level of probability or level of confidence

Steps in Inferential Statistics

3. Computing Calculated Value

- Use statistical test to derive some calculated value (e.g., t value or F value)

4. Obtain Critical Value

- a criterion used based on df and alpha level (.05 or .01) is compared to the calculated value to determine if findings are significant and therefore reject H_0

Steps in Inferential Statistics

5. Reject or Fail to Reject H_0

- CALCULATED value is compared to the CRITICAL value to determine if the difference is significant enough to reject H_0 at the predetermined level of significance
 - If CRITICAL value $>$ CALCULATED value \rightarrow fail to reject H_0
 - If CRITICAL value $<$ CALCULATED value \rightarrow reject H_0
 - If reject H_0 , *only supports H_1 ; it does not prove H_1*

Testing Hypothesis

- If reject H_0 and conclude groups are really different, it doesn't mean they're different for the reason you hypothesized
 - may be other reason
- Since H_0 testing is based on sample means, not population means, there is a possibility of making an error or wrong decision in rejecting or failing to reject H_0
 - Type I error
 - Type II error

Testing Hypothesis

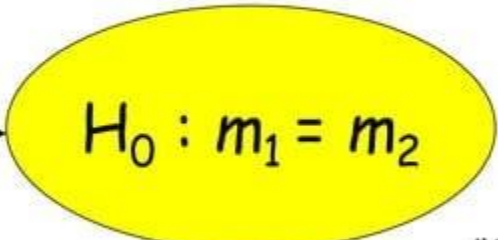
- Type I error -- rejecting H_0 when it was true (it should have been accepted)
 - equal to alpha
 - if $\alpha = .05$, then there's a 5% chance of Type I error
- Type II error -- accepting H_0 when it should have been rejected
 - If increase α , you will decrease the chance of Type II error

Inferential Statistics: uses sample data to evaluate the credibility of a hypothesis about a population



NULL Hypothesis:

NULL (*nullus* - latin): "not any" → no differences between means


$$H_0 : m_1 = m_2$$

Always testing the null hypothesis


"H- Naught"

Inferential statistics: uses sample data to evaluate the credibility of a hypothesis about a population



Hypothesis: Scientific or alternative hypothesis

Predicts that there are differences between the groups



$H_1 : m_1 \neq m_2$

Hypothesis

A statement about what findings are expected

null hypothesis

"the two groups will not differ"

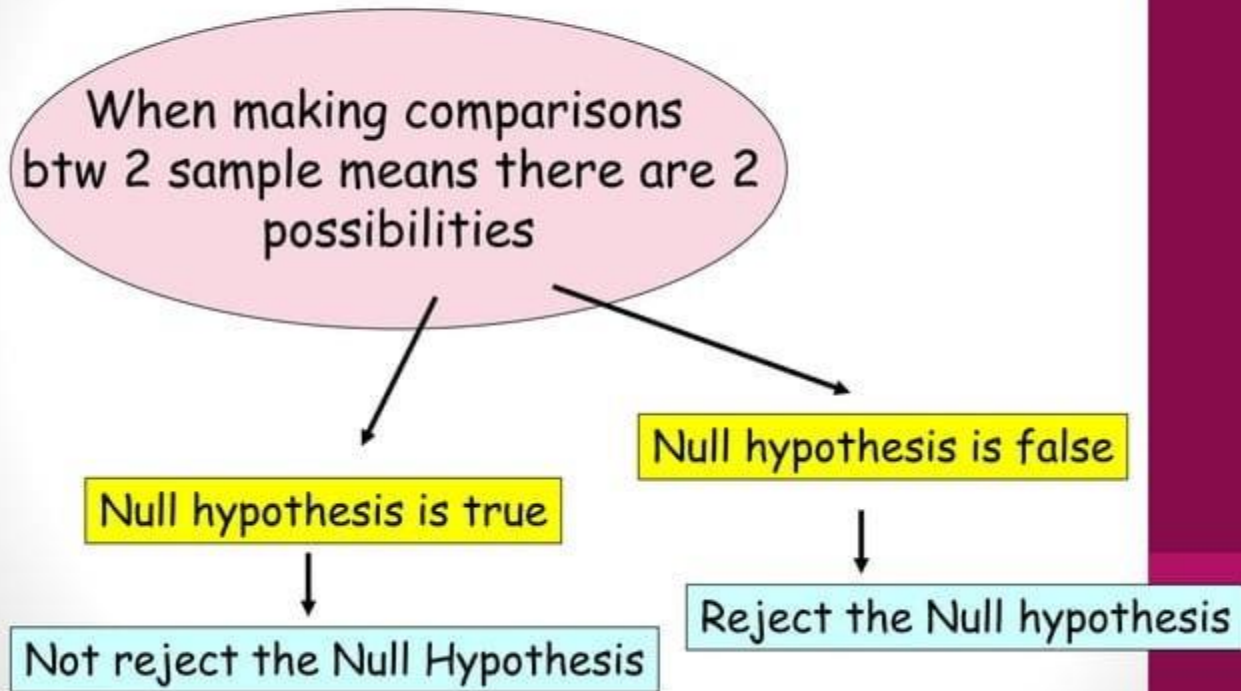
alternative hypothesis

"group A will do better than group B"

"group A and B will not perform the same"



Inferential Statistics



Possible Outcomes in Hypothesis Testing (Decision)

	Null is True	Null is False
Accept	Correct Decision	Error Type II Error
Reject	Error Type I Error	Correct Decision

Type I Error: Rejecting a True Hypothesis

Type II Error: Accepting a False Hypothesis

Hypothesis Testing - Decision

Decision Right or Wrong?

But we can know the probability of being right or wrong

Can specify and control the probability of making TYPE I or TYPE II Error

Try to keep it small...



ALPHA

the probability of making a type I error \rightarrow depends on the criterion you use to accept or reject the null hypothesis = significance level (smaller you make alpha, the less likely you are to commit error) 0.05 (5 chances in 100 that the difference observed was really due to sampling error - 5% of the time a type I error will occur)

Possible Outcomes in Hypothesis Testing

Alpha (α)

Difference observed is really just sampling error
The prob. of type one error

	Null is True	Null is False
Accept	Correct Decision	Error Type II Error
Reject	Error Type I Error	Correct Decision

When we do statistical analysis... if alpha
(p value- significance level) greater than 0.05



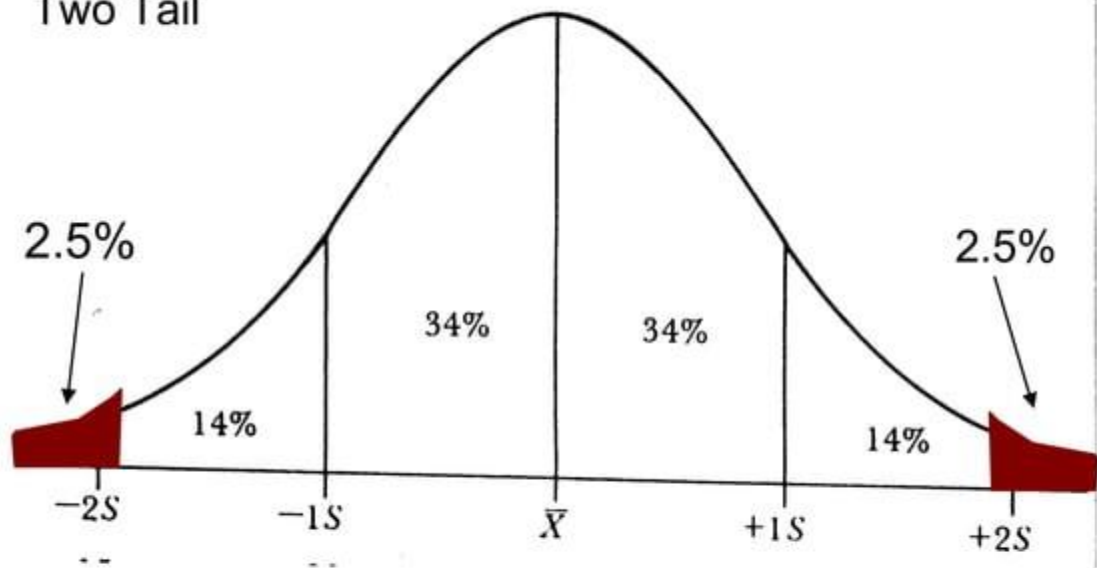
WE ACCEPT THE NULL HYPOTHESIS

is equal to or less than 0.05 we



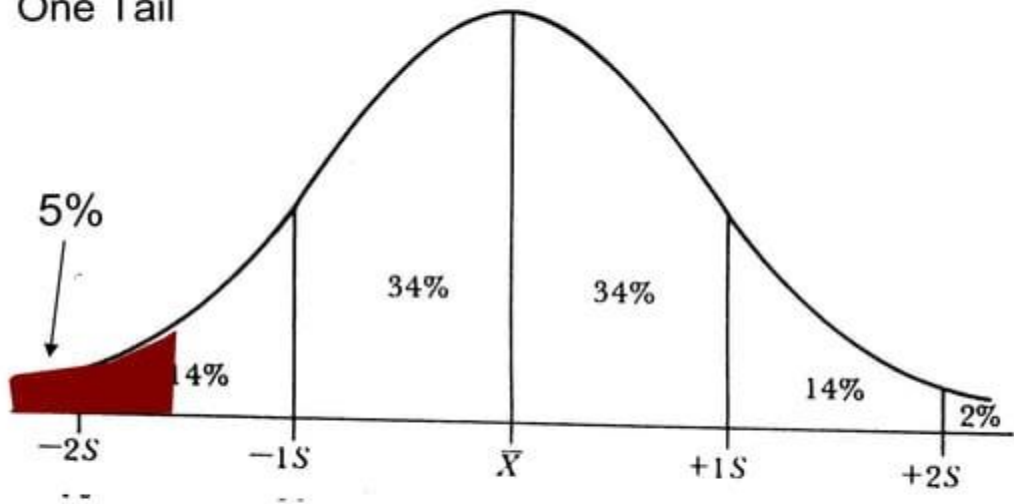
REJECT THE NULL (difference btw means)

Two Tail



5% region of rejection of null hypothesis
Non directional

One Tail



5% region of rejection of null hypothesis
Directional

BETA

Probability of making type II error \rightarrow occurs when we fail to reject the Null when we should have

Possible Outcomes in Hypothesis Testing

Beta (β)

	Null is True	Null is False
Accept	Correct Decision	Error Type II Error
Reject	Error Type I Error	Correct Decision

Difference observed is real
Failed to reject the Null

POWER: ability to reduce type II error

POWER: ability to reduce type II error
(1-Beta) - Power Analysis

The power to find an effect if an effect is present

1. Increase our n
2. Decrease variability
3. More precise measurements

Effect Size: measure of the size of the difference between means attributed to the treatment

Inferential statistics

Significance testing:

Practical vs statistical significance



Inferential statistics
Used for Testing for Mean Differences

T-test: when experiments include only 2 groups

- a. Independent
- b. Correlated
 - i. Within-subjects
 - ii. Matched

Based on the t statistic (critical values) based on
df & alpha level

Inferential statistics
Used for Testing for Mean Differences

Analysis of Variance (ANOVA): used when comparing more than 2 groups

1. Between Subjects
2. Within Subjects - repeated measures

Based on the f statistic (critical values) based on df & alpha level

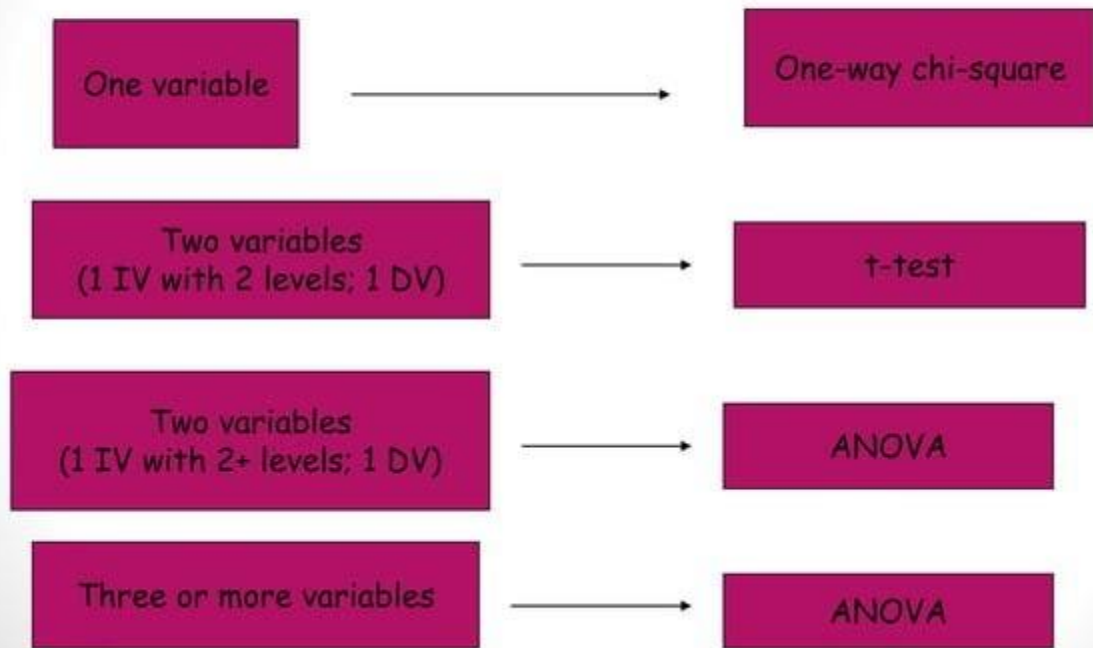
More than one IV = factorial (iv=factors)
Only one IV=one-way anova

Inferential statistics

Meta-Analysis:

Allows for statistical averaging of results
From independent studies of the same
phenomenon

Identifying the Appropriate Statistical Test of Difference



MODE

Book referred : BASIC STATISTICS FOR ECONOMISTS, V.K Global Publications

MODE

- Mode is the value which has the **greatest frequency in a distribution.**
- In simpler words, the value which **occurs most frequently in a series** is termed as **MODE**.
- It is another important measure of central tendency.
- Mode is denoted by the symbol '**Z**'.

Calculation of Mode

Individual Series

Inspection Method

By changing individual to discrete series

Discrete Series

Inspection Method

Grouping Method

Continuous Series

1. Inspection Method:-

This is the method which implies inspection (observation) of the items in the series. It is done by identifying the value that occurs most frequently in a series. Such value is called MODE.

Example - In a series: 20, 21, 23, 23, 23, 23, 25, 26, 26; the mode would be 23, since this value occurs most frequently than any other values in the series.

2. By changing the individual series into discrete series:-

INDIVIDUAL SERIES

When the numbers of items in a series is very large, individual series is **first converted into discrete series**. Then the **value corresponding to which there is highest frequency is identified**. Such value is called MODE.

Example - In a series:- 11.1, 10.9, 10.7, 11.1, 10.6, 10.7, 10.6, 10.9, 10.6, 10.5, 10.4, 10.6

We convert the series into a discrete series in ascending order-

SIZE:	10.4	10.5	10.6	10.7	10.9	11.1	11.3
FREQUENCY	1	1	5	2	2	2	1

10.6 is the Mode value since it appears maximum times in the given series.

1. Inspection Method:-

This is the method which implies inspection (observation) of the items in the series. It is done by identifying the value that occurs most frequently in a series. Such value is called MODE.

Example -

SIZE:	10.4	10.5	10.6	10.7	10.9	11.1	11.3
FREQUENCY	1	1	5	2	2	2	1

10.6 is the Mode value since it appears maximum times in the given series.

1. Grouping Method:-

- Determined by preparing **2 tables**:
 1. Grouping Table
 2. Analysis Table

Example -

X	7	8	9	10	11	12	13	14	15	16	17
f	2	3	6	12	20	24	25	7	5	3	1

STEP 1:
Grouping Table

**DISCRETE
SERIES**

X	I (f)	II (1+2)	III (2+3)	IV (1+2+3)	V (2+3+4)	VI (3+4+5)
7	2					
		5				
8	3			11		
			9			
9	6				21	
		18				
10	12					38
			32			
11	20			56		
		44				
12	24				69	
			49			
13	25					56
		32				
14	7			37		
			12			
15	5				15	
		8				
16	3					9
			4			
17	1					

STEP 2:
Analysis Table

DISCRETE SERIES

	Size (X)										
Col. No.	7	8	9	10	11	12	13	14	15	16	17
I							√				
II					√	√					
III						√	√				
IV				√	√	√					
V					√	√	√				
VI						√	√	√			
Total				1	3	5	4	1			

12 is the Mode value since it appears maximum times

CONTINUOUS SERIES

STEP 1:

Identify modal group either by inspection or grouping method

STEP 2:

Apply formula

$$Z = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Z = Mode

l_1 = lower limit of the modal class

f_1 = frequency of the modal class

f_0 = frequency previous to the modal class

f_2 = frequency next to the modal class

i = difference between the class interval

IMPORTANT POINTS:

- If the first class is the modal class, then $f_0 = 0$
- If the last class is the modal class, then $f_2 = 0$
- If the mode is ill-defined, then we use the formula $Z = 3M - 2X$

Example -

CONTINUOUS SERIES

Wages (Rs)	0-5	5-10	10-15	15-20	20-25	25-30	30-35
No. of workers	3	7	15 f_0	30 f_1	20 f_2	10	5

- By Inspection method, modal group is 15-20

$$Z = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Z = Mode = ?

l_1 = lower limit of the modal class = 15

f_1 = frequency of the modal class = 30

f_0 = frequency previous to the modal class = 15

f_2 = frequency next to the modal class = 20

i = difference between the class interval = 05

CONTINUOUS SERIES

$Z = \text{Mode} = ?$

$l_1 = \text{lower limit of the modal class} = 15$

$f_1 = \text{frequency of the modal class} = 30$

$f_0 = \text{frequency previous to the modal class} = 15$

$f_2 = \text{frequency next to the modal class} = 20$

$i = \text{difference between the class interval} = 5$

$$Z = 15 + \left\{ \frac{30 - 15}{2(30) - 15 - 20} \right\} \times 5$$

$$= 15 + \left\{ \frac{15}{25} \right\} \times 5$$

$$= 15 + \frac{15}{5}$$

$$= 15 + 3$$

$$\underline{\underline{Z = 18}}$$

Example -

CUMULATIVE FREQUENCY SERIES

Marks Between	Number of Students
10 and 15	4
10 and 20	12
10 and 25	30
10 and 30	60
10 and 35	80
10 and 40	90
10 and 45	95
10 and 50	97

Formula for calculating mode-

$$Z = l_1 + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times (i)$$

But first we need to convert the cumulative frequency series into **simple frequency series**:

Marks Between	Number of Students
10 and 15	4
10 and 20	12
10 and 25	30
10 and 30	60
10 and 35	80
10 and 40	90
10 and 45	95
10 and 50	97

Marks	f
10 – 15	4
15 – 20	8
20 – 25	18
25 – 30	30
30 – 35	20
35 – 40	10
40 – 45	5
45 - 50	2

By inspection, the modal class is 25 - 30

- $Z = \text{Mode}$
- $l_1 = \text{lower limit of the modal class} = 25$
- $f_1 = \text{frequency of the modal class} = 30$
- $f_0 = \text{frequency previous to the modal class} = 18$
- $f_2 = \text{frequency next to the modal class} = 20$
- $i = \text{difference between the class interval} = 5$

Applying the formula,

$$Z = l_1 + \frac{(f_1 - f_0) / (2f_1 - f_0 - f_2)}{1} \times (i)$$

$$Z = 25 + \{30 - 18 / 2(30) - 18 - 20\} \times 5$$

$$= 25 + \{12 / 60 - 18 - 20\} \times 5$$

$$= 25 + 60 / 22$$

$$= 25 + 2.73$$

$$\underline{\underline{Z = 27.73}}$$

INCLUSIVE SERIES

Example -

CLASS	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59
FREQUENCY	3	5	10	20	12	6	3	1

Since this is an inclusive series, we convert it into an exclusive one before solving and computing the mode.

CLASS	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59
<i>f</i>	3	5	10	20	12	6	3	1

INCLUSIVE SERIES

CLASS	FREQUENCY
19.5 – 24.5	3
24.5 - 29.5	5
29.5 – 34.5	10 <i>f₀</i>
34.5 – 39.5	20 <i>f₁</i>
39.5 – 44.5	12 <i>f₂</i>
44.5 – 49.5	6
49.5 – 54.5	3
54.5 – 59.5	1

By inspection method, modal class is 34.5 – 39.5

Formula used would be the same as before.

$$Z = l_1 + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times I$$

$$\begin{aligned} Z &= 34.5 + \left(\frac{20 - 10}{2 \times 20 - 10 - 12} \right) \times 5 \\ &= 34.5 + \left(\frac{10}{40 - 22} \right) \times 5 \\ &= 34.5 + \left(\frac{50}{18} \right) \times 5 \\ &= 34.5 + 2.77 \end{aligned}$$

$$\underline{\underline{Z = 37.27}}$$

Example -

BI-MODAL SERIES

Marks	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
<i>No. of students</i>	4	6	20	32	33	17	8	2

By inspection method, it is difficult to conclude which is the modal class

BI-MODAL SERIES

STEP 1: Grouping Table

Marks	I (f)	II (1+2)	III (2+3)	IV (1+2+3)	V (2+3+4)	VI (3+4+5)		
10—20	4] 10] 26] 30] 58] 85		
20—30	6							
30—40	20] 52] 65] 82] 58
40—50	32							
50—60	33] 50] 25] 58] 27] 2		
60—70	17							
70—80	8] 10] 2] 2] 2
80—90	2							

BI-MODAL SERIES

STEP 2: Analysis Table

Col. No.	10—20	20—30	30—40	40—50	50—60	60—70	70—80	80—90
I					√			
II			√	√				
III				√	√			
IV				√	√	√		
V		√	√	√	√	√	√	
VI			√	√	√			
Total		1	3	5	5	2	1	

As the maximum frequency occurs twice, it is a bi-modal series.

Calculation of Mean and Median

Marks	f	M.V. (m)	$A = 55$ d	$d' = d/10$	fd'	$c.f.$
10—20	4	15	-40	-4	-16	4
20—30	6	25	-30	-3	-18	10
30—40	20	35	-20	-2	-40	30
40—50	32	45	-10	-1	-32	62
50—60	33	55A	0	0	0	95
60—70	17	65	+10	+1	17	112
70—80	8	75	+20	+2	16	120
80—90	2	85	+30	+3	6	122
	$N = 122$				$\Sigma fd' = -67$	

$$\begin{aligned}\bar{X} &= A + \frac{\Sigma fd'}{N} \times i \\ &= 55 - \frac{67}{122} \times 10 = 55 - \frac{670}{122} \\ &= 49.51\end{aligned}$$

Median item = Size of $\frac{N}{2}$ th item = $\frac{122}{2} = 61$ th item which lies in 40—50.

$$\begin{aligned}M &= l_1 + \frac{\frac{N}{2} - c.f.}{f} \times i = 40 + \frac{61 - 30}{32} \times 10 \\ M &= 40 + \frac{310}{32} = 40 + 9.69 = 49.69\end{aligned}$$

Thus, $\bar{X} = 49.51$, $M = 49.69$

Now, $Z = 3M - 2\bar{X}$

$$= 3(49.69) - 2(49.51)$$

$$\therefore Z = 149.07 - 99.02 = 50.05$$

BI-MODAL SERIES

STEP 3:

Apply formula:

$$Z = 3\text{Median} - 2\text{Mean}$$

Empirical Relation between Mean, Median and Mode

- $Z = 3M - 2X$
- $X - Z = 3(X - M)$
- $M = \frac{1}{3}(2X - Z)$
- $X = \frac{1}{2}(3M - Z)$

X= Mean
M= Median
Z= Mode



STATISTICAL MEASURES

Measures of Center and Variation

VOCABULARY

- Data with one variable is called **univariate**.
- **Measures of central tendency** are numbers that describe the middle of the data.
- **Mean**-average_
 - Add up the numbers and divide by how many there are
- **Median**-middle number
 - Put the numbers in numerical order and find the one in the middle, if there is an even number average the two middle numbers
- **Mode**-number that occurs most frequently



FINDING MEASURES OF CENTER

10, 14, 15, 16, 20, 10, 11, 15, 19, 20, 10

γ. Find the mean

1. Add up the numbers

$$10+14+15+16+20+10+11+15+19+20+10=160$$

- γ. Divide by the number of numbers

$$160/11=14.545$$

γ. Find the median

1. Put numbers in order

10, 10, 10, 11, 14, 15, 15, 16, 19, 20, 20

- γ. Find the middle number.

15

ε. Find the mode

1. Which number occurs the most?

10



WHICH MEASURES OF CENTER IS BEST?

There are times when one measure of central tendency will better describe the data.

- **Mean**- when the data is spread out and you want an average of the values
- **Median**- when the data contains outliers
 - Outliers are values that don't fit the rest of the data
- **Mode**- when the data is tightly clustered around one or two values



EXAMPLE

- A new internet company has 3 employees who are each paid \$300,000, ten who are paid \$100,000 and sixty who are paid \$50,000. Which measure of central tendency best represents the pay at this company?
 - Median-because there are only a few employees who make a large amount of money these are outliers
 - Mode-because the vast majority of the employees make the same amount of money



MEASURES OF VARIATION OR DISPERSION

- A numerical value associated with how spread out the data is
- **Range**- largest value – smallest value
 - This is ONE number
- **Variance (σ^2)/Standard Deviation (σ)**- measures how much the data values differ from the mean

$$\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}}$$



FINDING VARIANCE AND STANDARD DEVIATION

○ To find variance σ^2 :

1. Find the mean
2. Find the difference between each data value and the mean.
3. Square each difference
4. Add them up
5. Divide by the number of data values

○ Find Standard deviation σ :

1. Square root the variance



FINDING MEASURES OF VARIATION

3, 5, 7, 9, 11

- Find the range

- Largest-smallest
 - $11-3=8$

- Find the standard deviation

- Find variance

∨ Mean = $3+5+7+9+11 = 35/5 = 7$

1 $(3-7)^2 + (5-7)^2 + (7-7)^2 + (9-7)^2 + (11-7)^2 =$

1 $(-4)^2 + (-2)^2 + (0)^2 + (2)^2 + (4)^2 =$

1 $16+4+0+4+16=20+0+20=40$

+ $40/5=8= \sigma^2$ -variance

+ $\sqrt{8}=2.83= \sigma$ -standard deviation



Sampling Techniques & Samples Types





Outlines

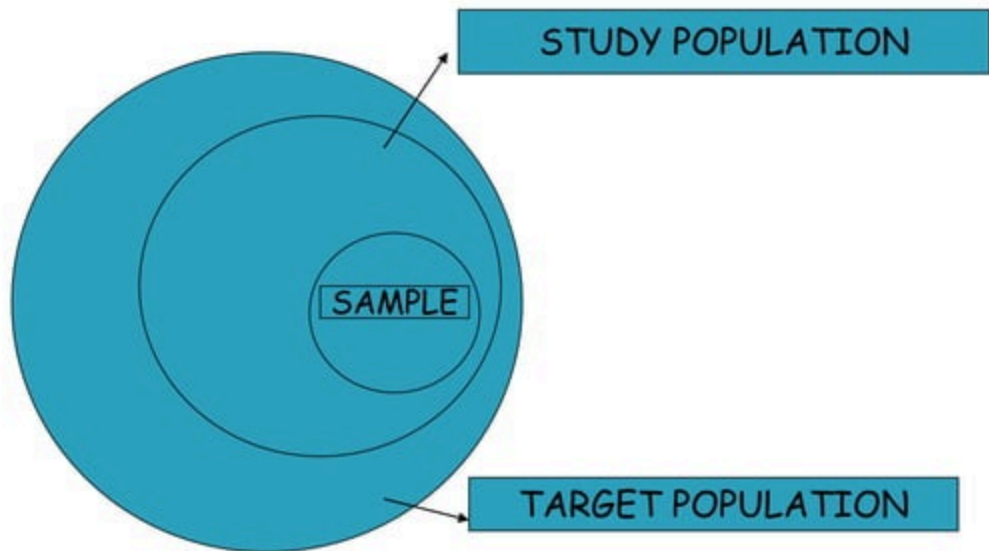
- Sample definition
- Purpose of sampling
- Stages in the selection of a sample
- Types of sampling in quantitative researches
- Types of sampling in qualitative researches
- Ethical Considerations in Data Collection

Sampling...

The process of selecting a number of individuals for a study in such a way that the individuals represent the larger group from which they were selected



SAMPLING.....



- A **sample** is “a smaller (but hopefully representative) collection of units from a population used to determine truths about that population” (Field, 2005)

- The **sampling frame**

A list of all elements or other units containing the elements in a population.

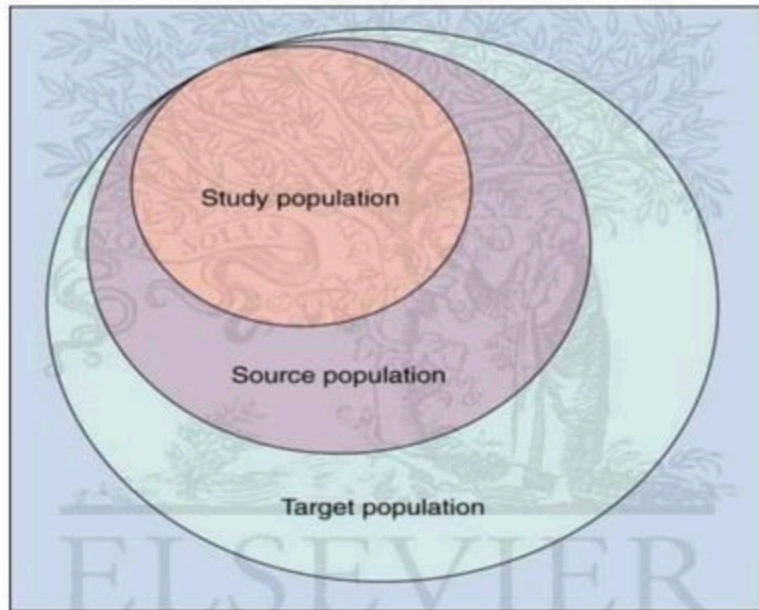
Population...

...the larger group from which individuals are selected to participate in a study

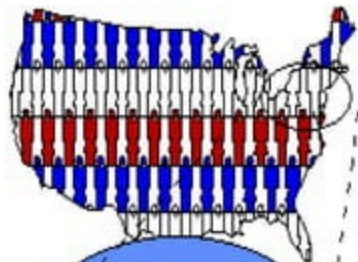


Target population

A set of elements larger than or different from the population sampled and to which the researcher would like to generalize study findings.



Who do you want to generalize to?



The Theoretical Population

What population can you get access to?



The Study Population

How can you get access to them?



The Sampling Frame

Who is in your study?



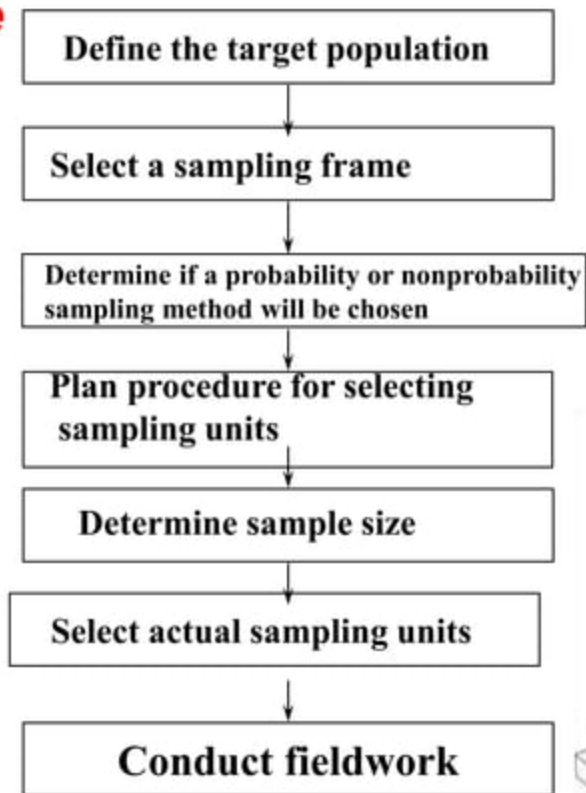
The Sample

The purpose of sampling...

- ▶ To gather data about the population in order to make an inference that can be generalized to the population



Stages in the Selection of a Sample



Quantitative Sampling

- ▶ Purpose – to identify participants from whom to seek some information
- ▶ Issues
 - Nature of the sample (random samples)
 - Size of the sample
 - Method of selecting the sample

Quantitative Sampling

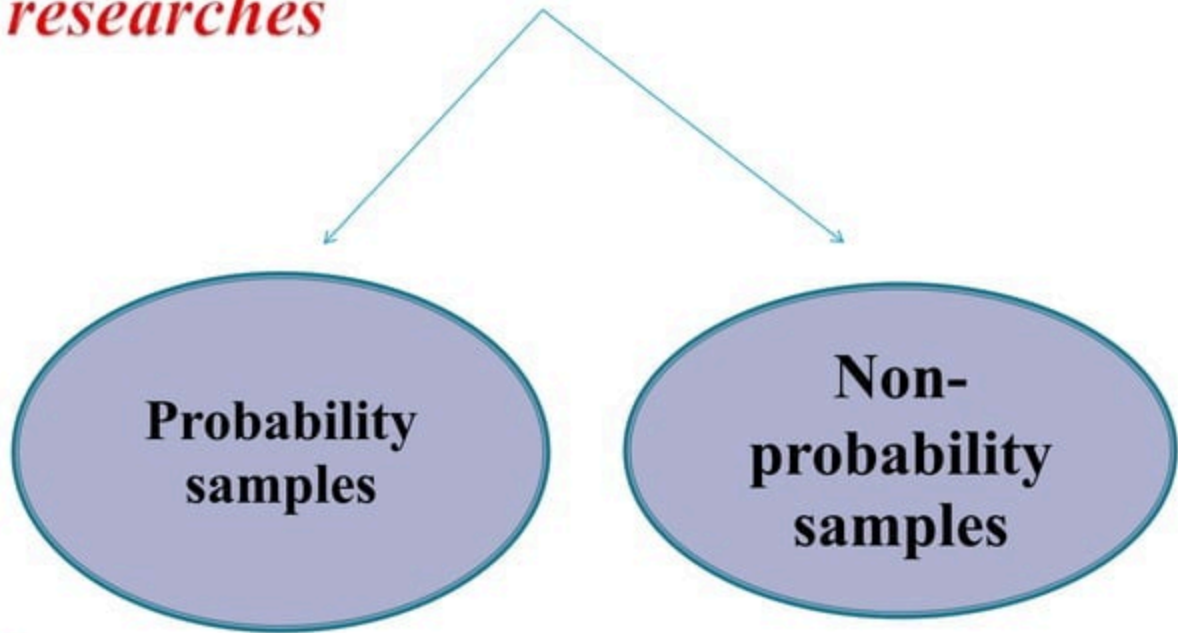
▶ Important issues

- Representation – the extent to which the sample is representative of the population
- Generalization – the extent to which the results of the study can be reasonably extended from the sample to the population
- Sampling error
The chance occurrence that a randomly selected sample is not representative of the population due to errors inherent in the sampling technique

Quantitative Sampling

- ▶ Important issues (continued)
 - Sampling bias
 - Some aspect of the researcher's sampling design creates bias in the data.
 - Three fundamental steps
 - Identify a population
 - Define the sample size
 - Select the sample

Types of sampling in quantitative researches



Selecting Random Samples

- ▶ Known as probability sampling
- ▶ Best method to achieve a representative sample
- ▶ Four techniques
 1. Random
 2. Stratified random
 3. Cluster
 4. Systematic

Selecting Random Samples

1. Random sampling

Selecting subjects so that all members of a population have an equal and independent chance of being selected

❖ Advantages

1. Easy to conduct
2. High probability of achieving a representative sample
3. Meets assumptions of many statistical procedures

❖ Disadvantages

1. Identification of all members of the population can be difficult
2. Contacting all members of the sample can be difficult

Selecting Random Samples

▶ Random sampling (continued)

- Selection process
 - Identify and define the population
 - Determine the desired sample size
 - List all members of the population
 - Assign all members on the list a consecutive number
 - Select an arbitrary starting point from a table of random numbers and read the appropriate number of digits



Selecting Random Samples

2. Stratified random sampling

- The population is divided into two or more groups called strata, according to some criterion, such as geographic location, grade level, age, or income, and subsamples are randomly selected from each strata.

Selecting Random Samples

▶ Stratified random sampling (continued)

◦ Advantages

- More accurate sample
- Can be used for both proportional and non-proportional samples
- Representation of subgroups in the sample

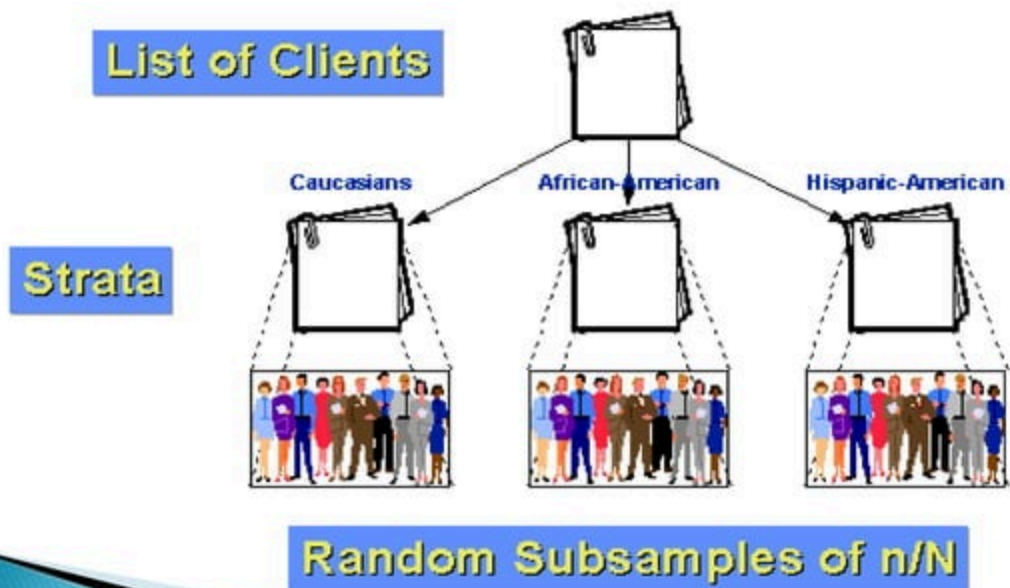
◦ Disadvantages

- Identification of all members of the population can be difficult
- Identifying members of all subgroups can be difficult

Selecting Random Samples

- ▶ Stratified random sampling (continued)
 - Selection process
 - Identify and define the population
 - Determine the desired sample size
 - Identify the variable and subgroups (i.e., strata) for which you want to guarantee appropriate representation
 - Classify all members of the population as members of one of the identified subgroups

Stratified random sampling



Selecting Random Samples

3. Cluster sampling

- ▶ The process of randomly selecting intact groups, not individuals, within the defined population sharing similar characteristics
- ▶ Clusters are locations within which an intact group of members of the population can be found
 - Examples
 - Neighborhoods
 - School districts
 - Schools
 - Classrooms

Selecting Random Samples

▶ Cluster sampling (continued)

◦ Advantages

- Very useful when populations are large and spread over a large geographic region
- Convenient and expedient
- Do not need the names of everyone in the population

◦ Disadvantages

- Representation is likely to become an issue

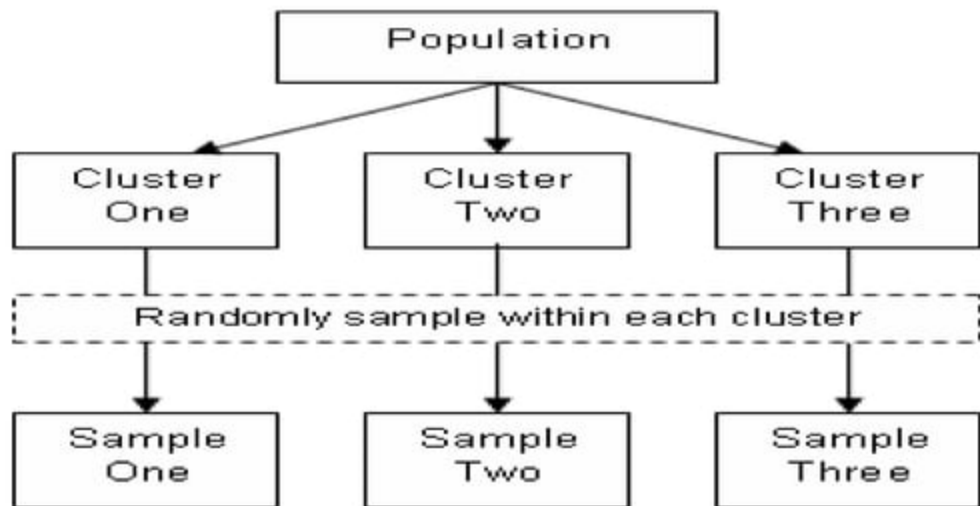
Selecting Random Samples

▶ Cluster sampling (continued)

◦ Selection process

- Identify and define the population
- Determine the desired sample size
- Identify and define a logical cluster
- List all clusters that make up the population of clusters
- Estimate the average number of population members per cluster
- Determine the number of clusters needed by dividing the sample size by the estimated size of a cluster
- Randomly select the needed numbers of clusters
- Include in the study all individuals in each selected cluster

Cluster sampling



Selecting Random Samples

4. Systematic sampling

- Selecting every K^{th} subject from a list of the members of the population
- Advantage
 - Very easily done
- Disadvantages
 - subgroups
 - Some members of the population don't have an equal chance of being included

Selecting Random Samples

▶ Systematic sampling (continued)

◦ Selection process

- Identify and define the population
- Determine the desired sample size
- Obtain a list of the population
- Determine what K is equal to by dividing the size of the population by the desired sample size
- Start at some random place in the population list
- Take every K^{th} individual on the list

Systematic sampling

- ▶ **Example**, to select a sample of 25 dorm rooms in your college dorm, makes a list of all the room numbers in the dorm. For example there are 100 rooms, divide the total number of rooms (100) by the number of rooms you want in the sample (25). The answer is 4. This means that you are going to select every fourth dorm room from the list. First of all, we have to determine the random starting point. This step can be done by picking any point on the table of random numbers, and read across or down until you come to a number between 1 and 4. This is your random starting point. For instance, your random starting point is "3". This means you select dorm room 3 as your first room, and then every fourth room down the list (3, 7, 11, 15, 19, etc.) until you have 25 rooms selected.

SAMPLE SIZE

- ▶ According to Uma Sekaran in Research Method for Business 4th Edition, Roscoe (1975) proposed the rules of thumb for determining sample size where sample size larger than 30 and less than 500 are appropriate for most research, and the minimum size of sample should be 30% of the population.
- ▶ The size of the sample depends on a number of factors and the researchers have to give the statistically information before they can get an answer. For example, these information like (confidence level, standard deviation, margin of error and population size) to determine the sample size.

Types of sampling in quantitative researches

Non-probability samples

(Random): allows a procedure governed by chance to select the sample; controls for sampling bias.



Nonrandom sampling methods...

1. Convenience sampling

2. Purposive sampling

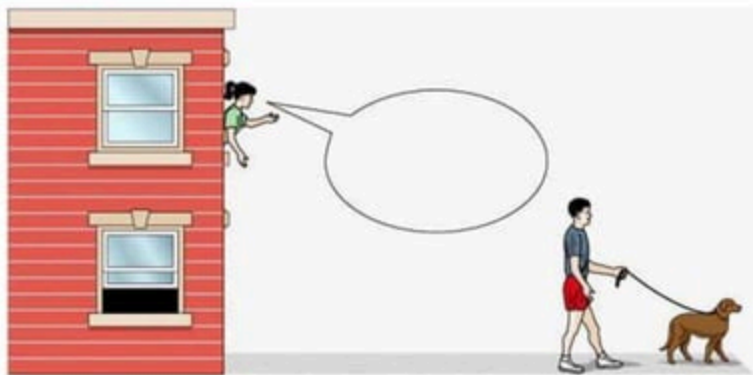
3. Quota sampling



1. Convenience sampling:

the process of including whoever happens to be available at the time

...called “accidental” or “haphazard” sampling



disadvantages...

...difficulty in determining how much of the effect (dependent variable) results from the cause (independent variable)

2. Purposive sampling:

the process whereby the researcher selects a sample based on experience or knowledge of the group to be sampled

...called “judgment” sampling



disadvantages...

...potential for inaccuracy in the researcher's
criteria and resulting sample selections

3. Quota sampling

the process whereby a researcher gathers data from individuals possessing identified characteristics and quotas



disadvantages...

...people who are less accessible (more difficult to contact, more reluctant to participate) are under-represented

Sampling in Qualitative Research



Sampling in Qualitative Research


Researchers in qualitative research select their participants according to their :

- 1) **characteristics**
- 2) **knowledge**



The purposeful sampling

It is when the researcher chooses persons or sites which provide specific knowledge about the topic of the study.




Types of Purposeful Sampling

- 1) Maximal Variation Sampling
- 2) Typical Sampling
- 3) Theory or Concept Sampling
- 4) Homogeneous Sampling
- 5) Critical Sampling
- 6) Opportunistic Sampling
- 7) Snowball Sampling



1 – Maximal Variation Sampling

It is when you select individuals that differ on a certain characteristic. In this strategy you should first identify the characteristic and then find individuals or sites which display that characteristic.




2- Typical Sampling

It is when you study a person or a site that is “typical” to those unfamiliar with the situation.

You can select a typical sample by collecting demographic data or survey data about all cases.


3-Theory or Concept Sampling

It is when you select individuals or sites because they can help you to generate a theory or specific concepts within the theory. In this strategy you need a full understanding of the concept or the theory expected to discover during the study.



4- Homogeneous Sampling

It is when you select certain sites or people because they possess similar characteristics. In this strategy, you need to identify the characteristics and find individuals or sites that possess it.




5- Critical Sampling

It is when you study an exceptional case represents the central phenomenon in dramatic terms.

6- Opportunistic Sampling

It is used after data collection begins, when you may find that you need to collect new information to answer your research questions.






Ethical Considerations in Data Collection

- It is the researcher's ethical responsibility to safeguard the story teller by maintaining the understood purpose of the research...
- The relationship should be based on trust between the researcher and participants.
- Inform participants of the purpose of the study.



- Being respectful of the research site, reciprocity, using ethical interview practices, maintaining privacy, and cooperating with participants.

 - Patton (2002) offered a checklist of general ethical issues to consider, such as:
 - ❖ reciprocity
 - ❖ assessment of risk
 - ❖ confidentiality,
 - ❖ informed consent
 - ❖ and data access and ownership.
- 

- Qualitative researchers must be aware of the potential for their own emotional turmoil in processing this information
- During the interview process, participants may disclose sensitive and potentially distressing information in the course of the interview..

